Leveraging Statistical Transliteration for Dictionary-Based English-Bengali CLIR of OCR'd Text

Utpal Garain[1] Arjun Das[1] David S. Doermann[2] Douglas D. Oard[2]
(1) INDIAN STATISTICAL INSTITUTE, 203, B. T. Road, Kolkata 700108, India.
(2) UNIVERSITY OF MARYLAND, College Park, MD USA
{utpal|arjundas}@isical.ac.in, {doermann|oard}@umd.edu

ABSTRACT

This paper describes experiments with transliteration of out-of-vocabulary English terms into Bengali to improve the effectiveness of English-Bengali Cross-Language Information Retrieval. We use a statistical translation model as a basis for transliteration, and present evaluation results on the FIRE 2011 RISOT Bengali test collection. Incorporating transliteration is shown to substantially and statistically significantly improve Mean Average Precision for both the text and OCR conditions. Learning a distortion model for OCR errors and then using that model to improve recall is also shown to yield a further substantial and statistically significant improvement for the OCR condition.

TITLE AND ABSTRACT IN BENGALI

OCR-কৃত নথি থেকে ইংরাজি-বাংলা অভিধান-ভিত্তিক CLIR-এর ক্ষেত্রে সংখ্যাতাত্ত্বিক লিপ্যন্তর-এর প্রভাব

সারাংশ

এই গবেষণাপত্রে অভিধান-বহির্ভূত ইংরাজি শব্দের বাংলায় লিপ্যন্তর বা প্রতিবর্ণীকরণ বিষয়ে পরীক্ষা এবং তার মাধ্যমে ইংরাজি-বাংলা CLIR-এর কার্যকারিতায় উন্নতিসাধন দেখানো হয়েছে। আমরা লিপ্যন্তর করার জন্য একটি সংখ্যাতাত্ত্বিক পদ্ধতি ব্যবহার করেছি এবং FIRE ২০১১ RISOT উপাত্ত বা ডাটা-এর সাহায্যে পদ্ধতিটির মূল্যায়ন করেছি। এটা দেখান হয়েছে যে এই লিপ্যন্তর-এর মাধ্যমে ইংরাজি-বাংলা CLIR-এর কার্যকারিতায় অনেকটাই উন্নতি পাওয়া যায়, যা পরিসংখ্যানগতভাবে উল্লেখযোগ্য। টাইপ অথবা OCR করা দু'ধরনের নথির ক্ষেত্রেই একই ধরনের ফল পাওয়া গেছে। পরিবর্তীকালে, OCR করা নথি থেকে CLIR করার ব্যাপারে OCR-এর ভুলগুলি থেকে একটি মডেল বানান হয়েছে ও দেখানো হয়েছে যে এই মডেল কিভাবে তথ্যোদ্ধারের ক্ষেত্রে আরও উন্নতি ঘটাতে পারে।

KEYWORDS: CLIR, OCR, English-Bengali, Dictionary based translation, Statistical transliteration, OCR error modeling, Stemming, Evaluation, FIRE-RISOT 2011.
KEYWORDS IN BENGALI: ভিন্ন ভাষায় তথ্যোদ্ধার, ছাপা অক্ষর/লেখা সনাক্তকরণ, ইংরাজি-বাংলা, অভিধান-ভিত্তিক অনুবাদ, সংখ্যাতাত্ত্বিক পদ্ধতিতে লিপ্যন্তর, OCR-এর ভুলের মডেলিং, স্টেমিং, মূল্যায়ন।

Proceedings of COLING 2012: Posters, pages 339–348,
COLING 2012, Mumbai, December 2012.

1 Introduction

Research in Cross-Language Information Retrieval (CLIR) has a long history, resulting in the formation of evaluation venues such as CLEF [CLEF, undated] and NTCIR [NTCIR, undated]. European and Oriental languages received the initial focus, but in recent years the CLEF evaluation has included Indian languages [Jagarlamudi, 2007]. Beginning in 2008, the Forum for Information Retrieval Evaluation (FIRE) [FIRE, undated] focused specifically on Indian languages. Monolingual Bengali retrieval was introduced to FIRE in 2008, and the first reported experiments with an English-to-Bengali (E2B) CLIR experiment design (i.e., English queries and Bengali documents) were reported in 2010, but the lack of translation resources for Bengali limited those experiments to simulation of CLIR using human query translation [Leveling, 2010]. This paper reports on the first fully automated experiments with E2B CLIR.

In case of E2B CLIR, the major challenge is limited Bengali resources. Although there is now an English-to-Bengali machine-readable dictionary available, we are not aware of any English-Bengali parallel corpus that is available for research use, any prior work (which is available for reuse) on English-Bengali transliteration, or any other lexical resources (e.g., multilingual WordNets) from which such a bilingual E2B translation lexicon might be extracted. We have therefore created an E2B lexicon of about 32,000 entries by manually cleaning the one available English-Bengali machine readable dictionary and we have trained a statistical transliteration tool to perform E2B translation.

A second important challenge with providing access to Bengali information is that a relatively large percentage of sources are only found in printed rather than digital form. In FIRE 2011, the RISOT track introduced a CLIR test collection (with both English and Bengali queries) for which two versions of a Bengali document collection are available: one containing digital Unicode text (text collection) and a second containing text recognized from document images using Optical Character Recognition (OCR collection) [Garain, 2011a]. Two groups reported results at FIRE 2011 on monolingual (Bengali-to-Bengali) OCR'd document retrieval [Garain, 2011b; Ghosh, 2011]. In this paper we report the first CLIR results for OCR'd Bengali documents using English queries, which to the best of our knowledge is only the second OCR-based CLIR results for any language (the first being English-to-Chinese [Tseng, 2001]). Our results show large and statistically significant improvements from statistical transliteration, statistical OCR error modeling, and their combination.

2 Statistical Transliteration for English-Bengali

To begin we used the transliteration method described by Virga and Khudanpur [Virga and Khudanpur, 2003]. In this method, transliteration is viewed as a simple character translation task. We used the Joshua open source statistical machine translation system [Li et al., 2009] which is reconfigured in [Irvine et al., 2010] for transliteration. Pairs of transliterated words and character-based n-gram language models are used in place of parallel sentences and word n-grams models. The Berkeley aligner [DeNero and Klein, 2007] is used to automatically align characters in pairs of transliterations. The language models are then trained on 2- through 10-gram sequences of target language characters. The goal is to minimize the edit distance between the system's output and the reference transliterations. This optimization is done by using the Joshua's Minimum Error Rate Training (MERT) and a character based BLEU score objective function (BLEU-4).

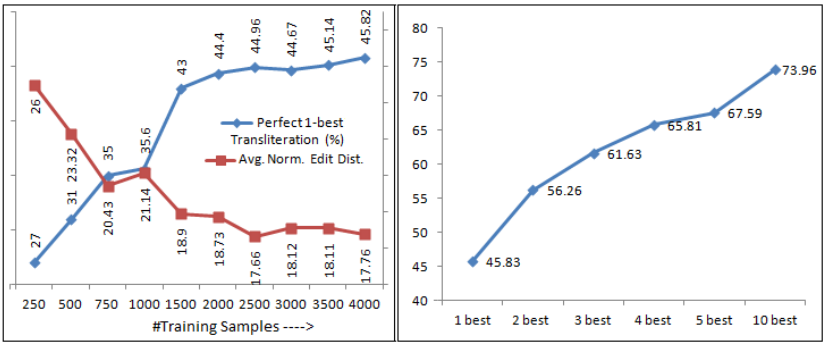

FIGURE 1 - Plots of (a) transliteration accuracy (1 best) and average normalized edit distance with the number of training samples and (b) N-best transliteration accuracy.

2.1 Training Data

For training, name pairs are mined from Wikipedia following an approach similar to one used by Irvine et al. [Irvine et al., 2010]. We obtained about 3,000 name pairs by considering the firstHeading field of the English and corresponding Bengali Wikipedia pages. Another 3,000 pairs were collected from other sources that contain both English and Bengali names of famous personalities, significant places (including names of Indian states, state capitals, important cities, etc.), movies, and other named entities. A Bengali language model was then built by first tagging the full Bengali news corpus from the FIRE test collection. This was done using the Stanford Part of Speech (POS) tagger, which was trained on approximately 8,000 tagged Bengali sentences (collected from Linguistic Data Consortium (LDC), University of Pennsylvania and the NLP Tool Contest at [ICON, 2009]). A total of ~30,000 unique named entities were identified through this process. The resulting named entities were then used to construct a character n-gram language model that includes n-grams up to length ten.

2.2 Evaluation of the English-Bengali Transliteration Model

For evaluating the transliteration module, our list of 6,000 name pairs was divided into 6 sets to facilitate a 6-fold cross validation. The ratio of training, development and test data for each fold was 4:1:1. Each set was used once as a test data and once as a development data. We report the Levenshtein edit distance, optionally normalized by the length of the reference string, and the F1 measure as intrinsic evaluation measures. As Figure 1(a) shows, increasing the number of training pairs yields substantial improvement between 250 to 1,500 pairs, with less dramatic improvements beyond 1500 training pairs - the system performance shows slower change as more data is added to the training set. For our final system (trained on about 6,000 pairs) the edit distance is 1.22, the normalized edit distance is 0.1776, and the F1measure is 0.7919. As Figure 1(b) shows, in about 46% of the cases, our system produced exactly the same string as the reference in the top position, increasing to about 74% of the cases when we look for an exact match somewhere in the top 10 candidates generated by our transliteration system. This suggests that using multiple transliteration alternatives in our CLIR system may be helpful.

3 English-Bengali CLIR System

In our CLIR model, the query in a source language (English) is first translated into the target language (Bengali) using an English-Bengali Bilingual dictionary. The out-of-vocabulary terms are transliterated. The query in the target language is then expanded using a generative stemmer (i.e., a system that generates terms that would stem to the same Bengali term). We conducted our CLIR (English query and Bengali collection) experiment both on clean and OCR'd collections separately. We refer to experiments on the clean collection as the "text condition" and the experiment on the OCR as the "OCR condition." For the OCR condition, the query terms were further expanded using an OCR error modeling technique.

3.1 English-Bengali Bilingual Dictionary

A bilingual dictionary is available from the Ankur project [Dictionary, undated], but as distributed it contains many unedited entries. We elected to retain only the edited entries, repeated entries were also automatically removed. This yielded 31,267 unique English terms. Most of the English terms have more than one Bengali translation. Only 14,764 English terms have only one Bengali meaning and others have multiple (up to 16) different translations. In total, there are 70,808 total term pairs (English term - Bengali translation). Although all English terms are one word, many of the Bengali translations are multiple word expressions. Out of 70,808 term pairs, for 26,915 cases the Bengali translation includes more than one word.

3.2 OCR Error Modeling

A key problem that distinguishes document image retrieval from other information retrieval problems is that character confusability during Optical Character Recognition (OCR) can result in mismatches between the (undistorted) query representation and the (distorted) document representation. For example consider an English query word "cat". Because of OCR errors "cat" may be distorted to "cot" if 'a' is misrecognized as 'o' in the OCR'd documents. Therefore, documents containing "cat" or "cot" or both should perhaps be retrieved for the query word "cat." One way of doing this is to expand the query (e.g., to include the word "cot" in the query in addition to "cat" whenever "cat" appears in the query posed by the user). In our case, we are using Bengali search terms. In order to do this well, the system needs some model for how Bengali characters are affected by OCR errors.

Our OCR error probabilities are built by comparing 20,000 documents containing 37 million characters of clean text with the electronic text generated from OCR. These pages are part of the RISOT collection on which we have tested our error model (note that the collection has about 63,000 documents). We used a dynamic programming approach to compare each pair of documents and to report statistics of Unicode errors. The report details which Unicode glyphs have been inserted, deleted, or substituted in the OCR text, and with what frequency each error was observed. The error counts for these 20,000 pages are combined and global statistics, referred to as "translation errors," are computed. From this knowledge we build a table (E_t) of triplets $<t_i, o_i, p_i>$ where t_i is translated to o_i with probability p_i, referred to as the corruption probability. Note that both t_i and o_i refer to a single codepoint or a group of codepoints. Our further investigation reveals that though the table contains more than 200 such triplets, the 75 top most frequent entries cover 80% of the error cases and our error model considers only them.

<top>
<num>26</num>
<title>সিঙ্গুরে জমি অধিগ্রহণ সমস্যা</title>
<desc>সিঙ্গুরে বামফ্রন্ট সরকারের জমি অধিগ্রহণ কর্মসূচি এবং ভূমি উচ্ছেদ প্রতিরোধ কমিটির বিক্ষোভ সংক্রান্ত নথি খুঁজে বার করো।</desc>
<narr>শিল্পোন্নয়নের জন্য সিঙ্গুরে কৃষি জমি অধিগ্রহণ, বামপন্থী ও বিরোধী দলের মধ্যে সঙ্ঘর্ষ, সাধারণ মানুষকে নিষ্ঠুর ভাবে হত্যা, সমাজের বিভিন্ন স্তরের মানুষের প্রতিবাদ ও সমালোচনা প্রাসঙ্গিক নথিতে থাকা উচিত।</narr>
</top>

<top>
<num>26</num>
<title>Singur land dispute</title>
<desc>The land acquisition policies of the Left Parties in Singur and the protest of Bhumi Ucched Protirodh Committee against this policy.</desc>
<narr>Relevant documents should contain information regarding the acquisition of agricultural land for industrial growth in Singur, the territorial battle between the Left Parties and the opposition parties, the brutal killing of the innocent people and the protests and the criticism by people from different sections of society.</narr>
</top>

FIGURE 2- Same topic in Bengali (left) and English (right).

3.3 Formation of the Translated Query

RISOT 2011 actually provided topics in only in Bengali, but the corresponding English topics are available from the FIRE 2010 E2B CLIR task. Fig. 2 is a sample topic in Bengali and English. We used Lemur toolkit for our experiments [Lemur, undated]. Following the Indri 5.1 query syntax, a title-only (T) query for the above topic would be posed as:

```
<query>
      <number>26</number>
      <text> #combine(singur land dispute)</text>
</query>
```

3.3.1 Dictionary-based Query Translation (DQT)

For a query in English, the basic idea is to look up each query word in the E2B lexicon, and for Out-of-Vocabulary (OOV) terms (i.e., those not found in the E2B lexicon) to use transliteration. For example, for the above query, "singur" (the name of a place) was not found in the E2B lexicon and thus was transliterated. For the term "land," 10 different translations are available in the E2B lexicon while the term "dispute" has 6 available translations. Since we don't have translation preference information available, the best known approach is to treat each alternate translation for a single term as members of a synonym set. In the query, these are combined using Indri's '#syn' operator [Pirkola, 1998]. We process these multiple word expressions (on the Bengali side of our E2B lexicon) as ordered phrases using Indri's '#1' proximity operator to enforce exact matching (e.g., #1(পৃথিবীর স্থলভাগ) will match only পৃথিবীর স্থলভাগ together and in that order). Before insertion of transliterations for OOV terms, the resulting Bengali query for the example shown above would be:

```
<query>
  <number>26</number>
    <text>
      #combine(#syn(#1(কৃষিভূমি) #1(পৃথিবীর স্থলভাগ) #1(জমিদারি) #1(অবতরণ
      করা) #1(জাহাজ) #1(গাড়ি থেকে নামা) #1(দেশ) #1(জমি) #1(স্থলবাহিনী) #1(জরিপ
      আমিন)) #syn(#1(তর্কাতর্কি) #1(প্রতিরোধ করা) #1(বিতর্ক) #1(প্রবল
      তর্ক) #1(সংঘাত) #1(বচসা করা)))
    </text>
</query>
```

3.3.2 Handling of OOVs

The English-Bengali transliteration module is used to generate one or more transliterated versions of each OOV term, returning the transliterations ranked in a best-first order. We then combine some number N of those transliterations, again using the #syn operator (as if they were alternative translations). When 10-best transliteration for the term "singur" in the above example is included, the Bengali query becomes:

```
<query>
  <number>26</number>
    <text>
      #combine(#syn(সিঙ্গুর সিনগুর সিনগর শিঙ্গুর সিংগুর সিংুর সীনগর সিনগুরে ঃসিঙ্গুর সিল্গুর)...)
    </text>
</query>
```

3.3.3 OCR Error Modeling (OEM)

Let $W_t = w_1 w_2 \ldots w_n$ be an n-codepoint query word. Note that we refer to codepoints (i.e., a single Unicode value) rather than characters to avoid confusion between the printed and digital representation; some Bengali glyphs are composed from more than one Unicode codepoint. We used the pruned set of 75 distortion probabilities learned in table E_t (see Section 3.2 above), treating all other Bengali code points as if they have zero distortion probability. Assuming that the codepoints of W_t are corrupted by OCR independently of each other, there may be many distorted versions of the word W_t. On average 27.5 variants are added for each term (minimum 0, maximum 128). We treat these distorted versions as synonyms, but this time we know the distortion probability and thus we use the Probabilistic Structured Query (PSQ) technique [Darwish and Oard, 2003], which is implemented by Indri's '#wsyn' operator. Let W_{ocr} be a possible distortion of query term W_{text}. We can then compute $P(W_{text} \mid W_{ocr})$ as

$$P(W_{text} \mid W_{ocr}) = \frac{P(W_{ocr} \mid W_{text}) P(W_{text})}{P(W_{ocr})}$$

where $P(W_{text})$ and $P(W_{ocr})$ are computed from the text and OCR collections. The term W_{ocr} is not considered in the expanded query if $P(W_{ocr}) = 0$. The third component, $P(W_{ocr}|W_{text})$ is basically $P(W_t \rightarrow W_s^i)$ which is computed from the error table E_t as discussed in Sec. 3.2.

4 Evaluation

The RISOT collection contains about 63,000 Bengali documents. We indexed both collections (Text and OCR) separately using the Lemur Toolkit and formed two types of queries: one from each topic's title field (T queries) and the other from each topic's title and description fields (TD queries). RISOT 2011 provides 92 topics for which one or more relevance judgments are available. We limited our evaluation to the 66 topics for which at least 5 relevant documents are known. Indri's default retrieval model [Ponte and Croft, 1998] is used.

4.1 Results

As a reference we report the monolingual MAP for the text condition using the original Bengali version of the topics. This yields 0.3205 for TD and 0.2649 for T queries (runs T1 and T6 in Table 1). When we perform CLIR without transliteration (the DQT technique alone), only 73%

of the query terms are found in the E2B lexicon. As a result, we get relatively poor results; a MAP of 0.1230 for TD queries and 0.0965 for T queries (runs T2 and T7). Translation ambiguity is not actually hurting us much in this case: manually selecting the best single-word Bengali translations from the alternatives available in the B2C lexicon (to eliminate both the '#1' and '#syn' operators) results in only small apparent improvements (runs T3 and T8) that are not statistically significant (runs T2:T3, T7:T8; $p>0.1$ by a two-tail t-test).

TABLE 1 – English-Bengali CLIR results for RISOT 2011 Collection, Text Condition

Run	Q	Retrieval Condition	Processing	MAP	MAP %	P@5	P@10	Rprec
T1	TD	Monolingual	--	0.3205	100%	0.3762	0.3182	0.3083
T2	TD	CLIR	DQT	0.1230	38%	0.1370	0.1167	0.1240
T3	TD	CLIR	DQT (Manual selection)	0.1269	40%	0.1665	0.1433	0.1410
T4	TD	CLIR	DQT + OOV	0.2645	83%	0.2887	0.2558	0.2605
T5	**TD**	**CLIR**	**DQT + OOV + Stemming**	**0.3306**	**103%**	**0.3609**	**0.3197**	**0.3204**
T6	T	Monolingual	---	0.2649	100%	0.3109	0.2630	0.2550
T7	T	CLIR	DQT	0.0965	36%	0.1114	0.1068	0.0980
T8	T	CLIR	DQT (Manual selection)	0.0969	37%	0.1271	0.1094	0.1080
T9	T	CLIR	DQT + OOV	0.2186	83%	0.2386	0.2114	0.2150
T10	**T**	**CLIR**	**DQT + OOV + Stemming**	**0.2689**	**102%**	**0.2935**	**0.2600**	**0.2648**

Incorporating the 10-best transliterations for OOV English query terms (with fully automatic E2B translation for all other English query terms) yields substantial and statistically significant improvement over DQT alone (runs T2:T4, T7:T9; $p<0.01$). Smaller values of N (not shown) do somewhat less well (MAP improvements from 1-best to 3-best, 3-best to 5-best and 5-best to 10-best are statistically significant at $p<0.05$), and larger values of N yield no further improvement.

As Bengali is a highly inflectional language, we then used a statistical stemmer [Paik et al., 2011]. Given a query term, it generates all possible variations of the words. The stemming yields a statistically significant improvements for both T and TD queries (runs T4:T5, T9:T10; $p<0.01$). The best CLIR results are thus obtained from combining dictionary based translation with transliteration of OOVs and generative stemming. Indeed, this combination achieved MAP values that slightly exceed those of monolingual retrieval (without stemming), demonstrating that the monolingual condition should be considered as a reference and not as an upper bound.

Table 2 shows comparable results for our experiments with the OCR condition. Again, DQT alone does relatively poorly (runs O2 and O8) and manual selection of single-word translations again does not yield a significant improvement (runs O2:O3, O8:O9; $p>0.1$). As with the text condition, transliteration yields significant improvements for the OCR condition (runs O2:O4, O8:O10; $p<0.01$). Further statistically significant improvement results from OCR error modeling (see Section 3.2.3) (runs O4:O5, O10:O11; $p<0.01$). Finally, the best overall results for the OCR condition resulted from combining transliteration of OOV terms, modeling of OCR errors, and stemming (runs O5:O6, O11:O12; $p<0.01$). For the OCR condition, this combination achieves MAP valued near, but below, the corresponding monolingual MAP for the text condition.

Note that stemmed monolingual retrieval yielded MAPs equal to 0.3929 (TD) and 0.3125 (T). If these MAPs are used as baselines, CLIR (text condition) best performance is only 84% (and 86% for T queries) of the best monolingual performance for TD queries and CLIR OCR condition MAPs are only 74% and 75% of the best monolingual results for TD and T queries.

TABLE 2 – English-Bengali CLIR results for RISOT 2011 OCR'd Collection
(The rows for Runs T1 and T6 are reference results from text condition)

Run	Q	Retrieval Condition	Processing	MAP	MAP%	P@5	P@10	Rprec
T1	TD	Mono+Text	--	0.3205	100%	0.3762	0.3182	0.3083
O1	TD	Monolingual	--	0.2689	84%	0.2420	0.2420	0.4166
O2	TD	CLIR	DQT	0.0813	25%	0.1025	0.0854	0.0679
O3	TD	CLIR	DQT (Manual selection)	0.0848	26%	0.1150	0.0938	0.0864
O4	TD	CLIR	DQT + OOV	0.1866	58%	0.2529	0.2063	0.1901
O5	TD	CLIR	DQT+OOV+OEM	0.2650	83%	0.3338	0.2723	0.2509
O6	**T**	**CLIR**	**DQT+OOV+OEM+Stem**	**0.2915**	**91%**	**0.3672**	**0.2996**	**0.2760**
T6	T	Mono+Text	--	0.2649	100%	0.3109	0.2630	0.2550
O7	T	Monolingual	---	0.2222	84%	0.2000	0.2000	0.3330
O8	T	CLIR	DQT	0.0672	25%	0.0847	0.0706	0.0560
O9	T	CLIR	DQT (Manual selection)	0.0701	26%	0.0950	0.0775	0.0710
O10	T	CLIR	DQT+OOV	0.1607	61%	0.1694	0.1494	0.1490
O11	T	CLIR	DQT + OOV + OEM	0.2121	80%	0.2236	0.1972	0.1965
O12	**T**	**CLIR**	**DQT+OOV+OEM+Stem**	**0.2333**	**88%**	**0.2460**	**0.2169**	**0.2162**

Conclusion and perspectives

We have described an English-to-Bengali CLIR system and showed that the basic dictionary-based method can be significantly improved by using transliteration to accommodate OOV terms. Our system has been evaluated using both a clean (digital) text and an OCR condition, and for the OCR condition modeling of OCR errors has also been shown to significantly improve retrieval effectiveness. Our reliance on affordable statistically trained techniques for stemming, transliteration, and OCR error modeling, suggests that similar techniques could reasonably be tried with any language for which a moderately large bilingual dictionary (and a suitable test collection) are available.

Several significant resources are resulted in from this research. A list of 6,000 English-Bengali proper names has been generated. An English-Bengali transliteration system is now available (the system can easily be modified to a B2E transliteration system). The English-Bengali cleaned dictionary consisting of about 32,000 entries is another sharable resource which is generated under this work. All these resources are made freely available for doing further research in NLP and CLIR involving Bengali. Comparison with stemmed monolingual retrieval suggests that further improvements might be possible in some cases where our present E2B lexicon has gaps. In these cases, our present transliteration system fails to find the correct transliteration. This suggests that continued work on tuning and robustness might be productive. As next steps, we plan to try (i) pre-translation and post-translation blind relevance feedback to improve robustness and (ii) mining comparable corpora to learn additional translation candidates as an additional way of filling lexical gaps.

Acknowledgments

One of the authors thanks the Indo-US Science and Technology Forum for providing him with a support to conduct a part of this research at the University of Maryland. Thanks to Ann Irvine of John Hopkins University and Jiaul Paik of ISI, Kolkata for their kind help.

References

CLEF: The Cross-Language Evaluation Forum. http://clef-campaign.org.

Darwish, K. and Oard, D. (2003). Probabilistic Structured Query Methods, In ACM-SIGIR, pages 338-344.

DeNero, J. and Klein, D. (2007). Tailoring word alignments to syntactic machine translation, In ACL, pages 17-24.

Dictionary: http://www.bengalinux.org/english-to-bengali-dictionary/

FIRE: The Forum for Information Retrieval Evaluation. http://www.isical.ac.in/~clia/

Garain, U., Paik, J., Pal, T., Majumder, P., Doermann, D. and Oard, D. (2011a). Overview of FIRE 2011 RISOT Task, In Forum for Information Retrieval Evaluation (FIRE), Mumbai, India.

Garain, U., Doermann, D. and Oard, D. (2011b). Maryland at FIRE 2011: Retrieval of OCR'd Bengali, In Forum for Information Retrieval Evaluation (FIRE) 2011, Mumbai, India.

Ghosh, K. and Parui, S.K. (2011). Retrieval from OCR text: RISOT track, In Forum for Information Retrieval Evaluation (FIRE) 2011, Mumbai, India.

ICON (2009). NLP Tool Contest: Parsing, In 7th International Conference on Natural Language Processing, Hyderabad, India.

Irvine, A., Callison-Burch, C. and Klementiev, A. (2010). Transliterating From All Languages, In AMTA 2010, Denver, Colorado.

Jagarlamudi, J. and Kumaran, A. (2007). Cross-lingual Information Retrieval System for Indian Languages, In CLEF Workshop.

Lemur: http://www.lemurproject.org/indri/

Leveling, J., Ganguly, D. and Jones, G.J.F. (2010). DCU@FIRE2010: Term Conflation, Blind Relevance Feedback, and Cross-Language IR with Manual and Automatic Query Translation, In Forum for Information Retrieval Evaluation (FIRE) 2010, Ghandhinagar, India.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L. Thornton, W., Weese, J. and Zaidan, O. (2009). Joshua: An open source toolkit for parsing based machine translation, In EACL 4th Workshop on Statistical Machine Translation, Athens, Greece.

NTCIR: fttp://research.nii.ac.jp/ntcir.

Paik, J., Mitra, M., Parui, S., and Järvelin, K. (2011). GRAS: An Effective and Efficient Stemming Algorithm for Information Retrieval, In ACM Transactions on Information Systems, 29(4).

Pirkola, A. (1998). The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In ACM SIGIR, Pages 55-63.

Ponte, J.M. and Croft, W.B. (1998). A language modeling approach to information retrieval, In ACM SIGIR, Pages 275-281.

Tseng, Y.-H. and Oard, D.W. (2001). Document Image Retrieval Techniques for Chinese, In Symposium on Document Image Understanding Technology, pages 151-158, Columbia, MD.

Virga, P. and Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Lingual Information Retrieval, In ACL Workshop on Multi-lingual Named Entity Recognition Combining Statistical and Symbolic Models, Sapporo , Japan.